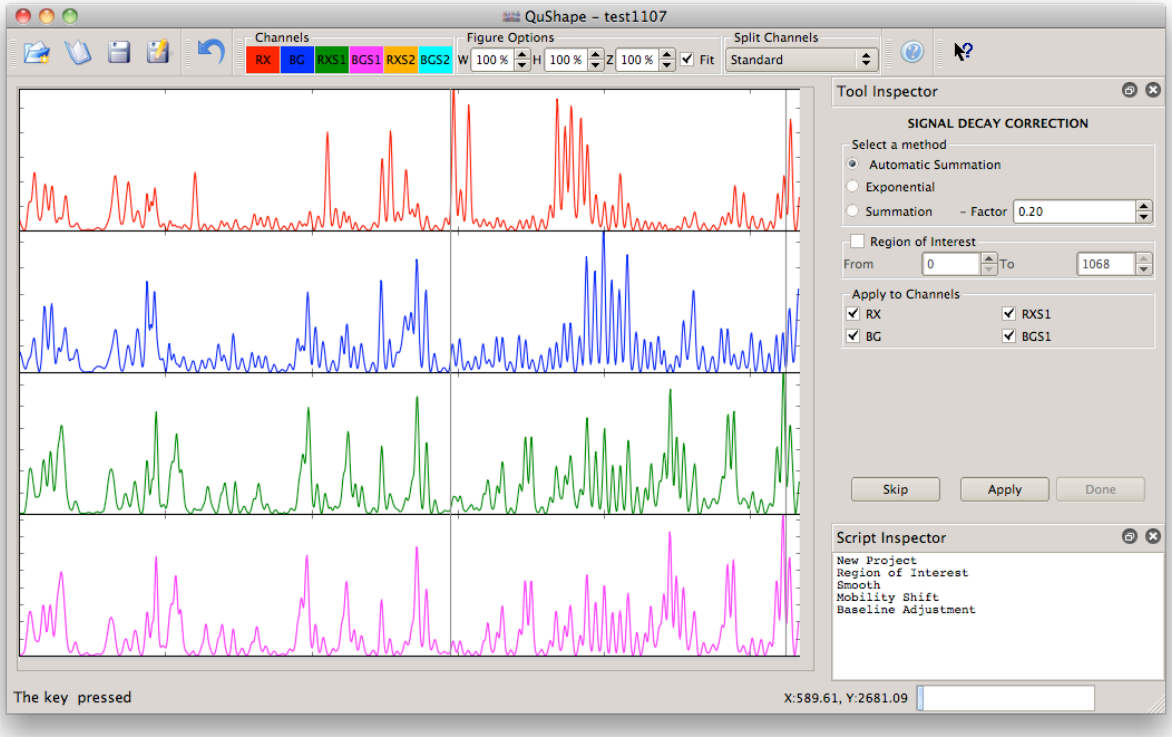
SHAPE DATA ANALYSIS USING QUSHAPE

1. INTRODUCTION

QuShape is a comprehensive user-friendly software package designed to perform fully automated analysis of SHAPE experimental data. QuShape requires no other software to fully analyze raw experimental capillary electrophoresis data. QuShape can be run in an automatic mode with default analytical procedures but contains an array of alternative algorithmic procedures and parameter controls that the user can apply if not satisfied with the default results.

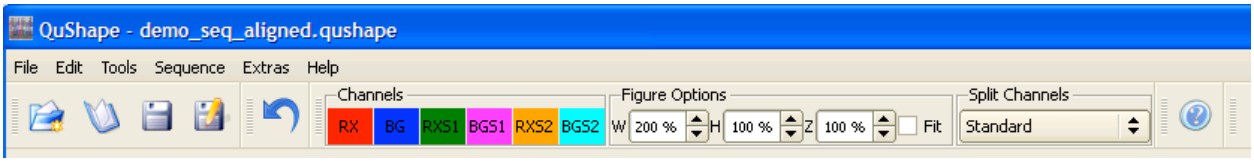
2. GRAPHIC USER INTERFACE

The user controls QuShape via a graphic interface. This interface includes the main **Data View** window, the **Tool Inspector** window, and the **Script Inspector** window. Results of every operation are plotted in the **Data View** window, allowing the user to monitor the quality of each data processing step. The user can vary the graphic display format via a set of control buttons. At each data processing step, the **Tool Inspector** window offers the user additional analytical tools that can be employed if the user is not satisfied with the results of the automatic procedure.

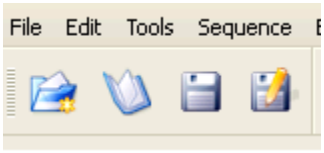


2.1. MAIN MENU AND TOOLS BAR

Four groups of icons are found at the top of the QuShape screen: **File-handling**, **Channels**, **Figure Options**, and **Split Channels**.



File-handling icons: These icons allow users to create or save projects.



New Project – Click this icon to create a new project.

Open Project – Click this icon to open an existing project.

Save Project – Click this icon to save the project.

Save Project As – Click this icon to save the project under a different name.

Channels icons: Six differently colored label widgets relate colors of the lines in the **Data View** window to specific data channels (these labels refer to the two-capillary approach used to resolve the chemical probing experiment data).

RX (red): (+) reagent reaction signal.

BG (blue): (–) reagent reaction signal.

RXS1 (green) and **RXS2** (orange): Dideoxynucleotide sequencing signals in the (+) reagent capillary.

BGS1 (magenta) and **BGS2** (cyan): Dideoxynucleotide signals in the (–) reagent capillary.

Figure Options icons: Four icons control the display scales of the **Data View** window.

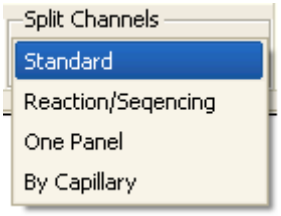
W (width): Horizontal scale. Uncheck **Fit** button to enable this and the next two controls.

H (height): Vertical scale of the entire window.

Z (zoom): Vertical scale within each panel.

Fit: Checking this box will fit the entirety of each plot to the display area.

Split Channels: Four options control the number of display panels in the **Data View** window and the data signals that are shown in each.



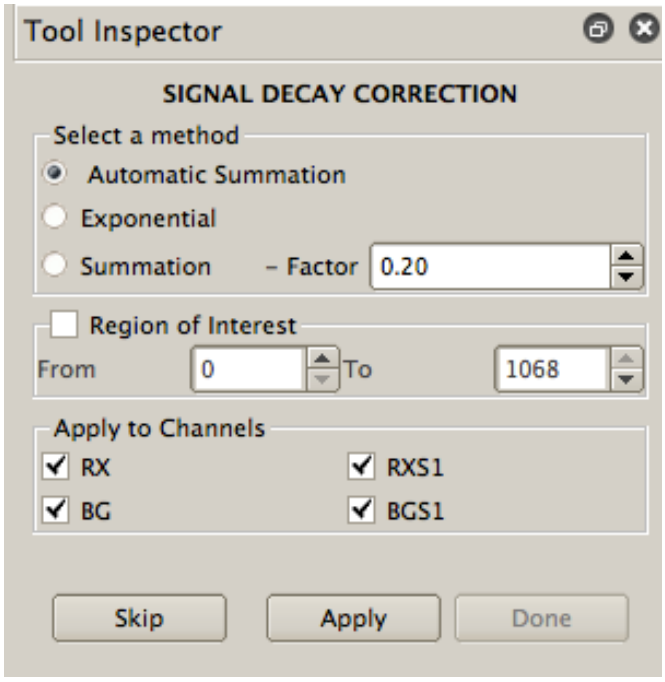
Standard: Each data channel is shown in a separate panel.

Reaction/Sequencing: The reactions channels (RX and BG) are paired together in the top panel and the sequencing channels (RXS, BGS) are paired together in the bottom panel. This pairing is particularly useful in the Signal Alignment step.

Capillary: The reaction and sequencing signals measured in the (+) reagent capillary (RX and RXS) are paired together in the top panel and the reaction, and sequencing signals measured in the (–) reagent capillary (BG, BGS) are paired together in the bottom panel. This pairing is particularly useful in the Mobility Shift step.

One Panel: All data are superimposed in a single plot. If used, RXS2 is plotted with RXS1 in the same panel and BGS2 is plotted with BGS1 in the same panel.

2.2. TOOL INSPECTOR



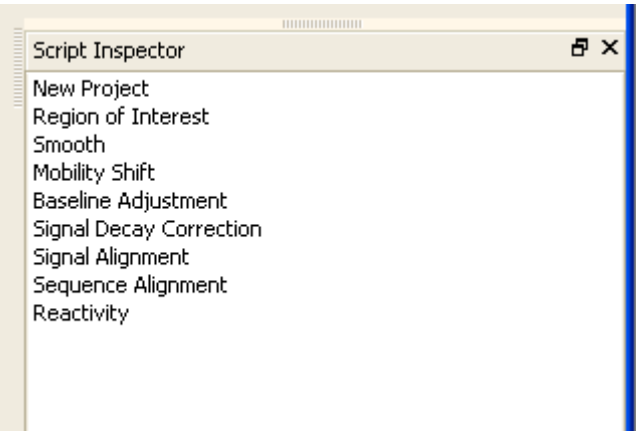
Data processing steps are executed on command from the **Tool Inspector** window. Selection of each successive data processing step is either automatic (default option) or is performed by clicking on a particular tool from among those listed in the **Tools** and **Sequence** pull-down menu options on the top menu bar. The selection of tools is enabled once a data file has been loaded.

The **Tool Inspector** window has three standard buttons, common to all tools, as well as unique control buttons specific to particular tools. The three standard buttons are **Apply**, **Done**, and **Skip**. The **Apply** button initiates the execution of the tool. Click this button after selecting appropriate settings for the tool (if different from the default settings). If the result of the tool execution is satisfactory (ascertained by checking the display in the **Data View** window), click the **Done** button to accept the result. The name of this tool will then appear in the **Script Inspector** window (registering it as having been executed), and the next appropriate tool will be displayed in the **Tool Inspector** window ready for action. If the current tool

is not needed, click the **Skip** button to cancel any modifications this tool may have made to the loaded data and move the program to the next appropriate tool.

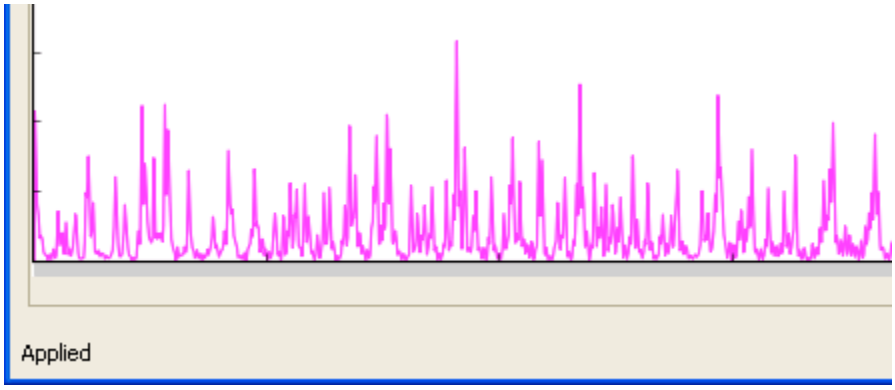
Some tools have an **Apply to Channels** option. The tool will be applied only to those channels with boxes checked. In addition, some tools have the option of being applied to a subset of the data specified by an elution time interval using the **Region of Interest** spin-box widgets.

2.3. SCRIPT INSPECTOR



The **Script Inspector** displays the sequence of tools applied to the active dataset. Using this window, it is possible to review the effects of the tools used and to restore an earlier state and continue processing. By double-clicking a particular tool in the script, the final product of this tool will be shown in the **Data View** window. In this way, it is possible to observe the intermediate results of each applied tool.

2.4. STATUS BAR



Some of the tools are computationally intensive, and the processes take tens of seconds to execute. During their execution (after pressing the **Apply** button), the left-bottom corner of the screen will display the "**Applying…**" message. Once the operation is finished, this message will change to "**Applied**", and the **Done** button in the **Tool Inspector** window will become enabled (its appearance will change from dim to sharp contrast).

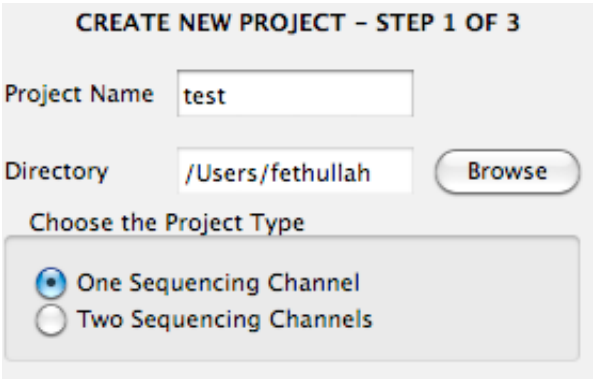
3. NEW PROJECT

There are three ways to create a new project to analyze SHAPE experimental data:

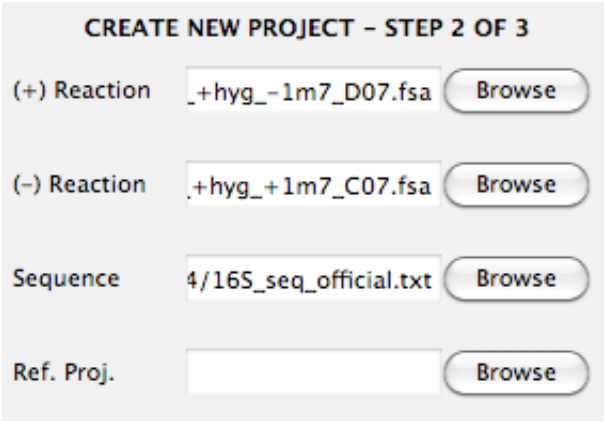
(1) Click **New Project** from the **File** menu

(2) Click  icon in the toolbar

(3) Press Ctrl-N



A wizard will appear to guide project set up, which involves three steps. In the first step, enter the name of the project in the Project Name box and choose the working directory by browsing. Next, select the project type. If there is just one sequencing lane in each capillary (RXS1 or BGS1) in the files obtained from electrophoresis, select **One Sequencing Channel**. Otherwise select the second option, **Two Sequencing Channels**. With two sequencing channels per capillary (RXS1 and RXS2, BGS1 and BGS2), sequence alignment will be performed more accurately. Press the **Next** button to go to next step.

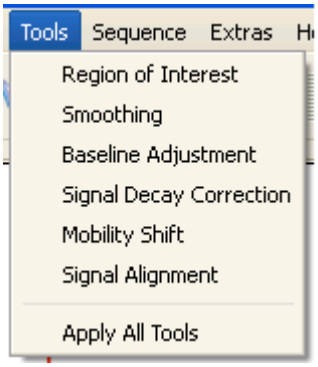


In the second step, select the data files using the **Browse** button. Text or ABIF formatted **(+)** and **(–) Reaction** files are both acceptable. The RNA **Sequence** file (.txt, .seq, .fasta) or a **Reference Project** file (.qushape) can be selected in the same way. Click the **Next** button to go to the last step.

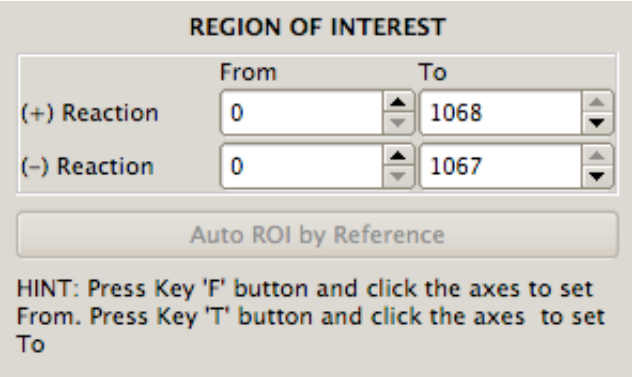
In the final step, the channels are selected in the files. Select channels in the (+) reaction file to specify RX and RXS1. For the sequencing ladder, the ddNTP type (ddC, ddG, ddT, ddA) must be selected in the same way that BG and BGS1 are selected in the (–) Reaction file. If there is another sequencing lane, RXS2 and BGS2 must be selected in the (+) and (–) Reaction files, respectively. After specifying all the channels, press the **Apply** button to view the data display in the **Data View** window. If all selections are correct, press the **Done** button to proceed to the analysis. If there is a problem with specified options, use the **Back** button to go to the previous dialog to change the parameters.

4. TOOLS MENU

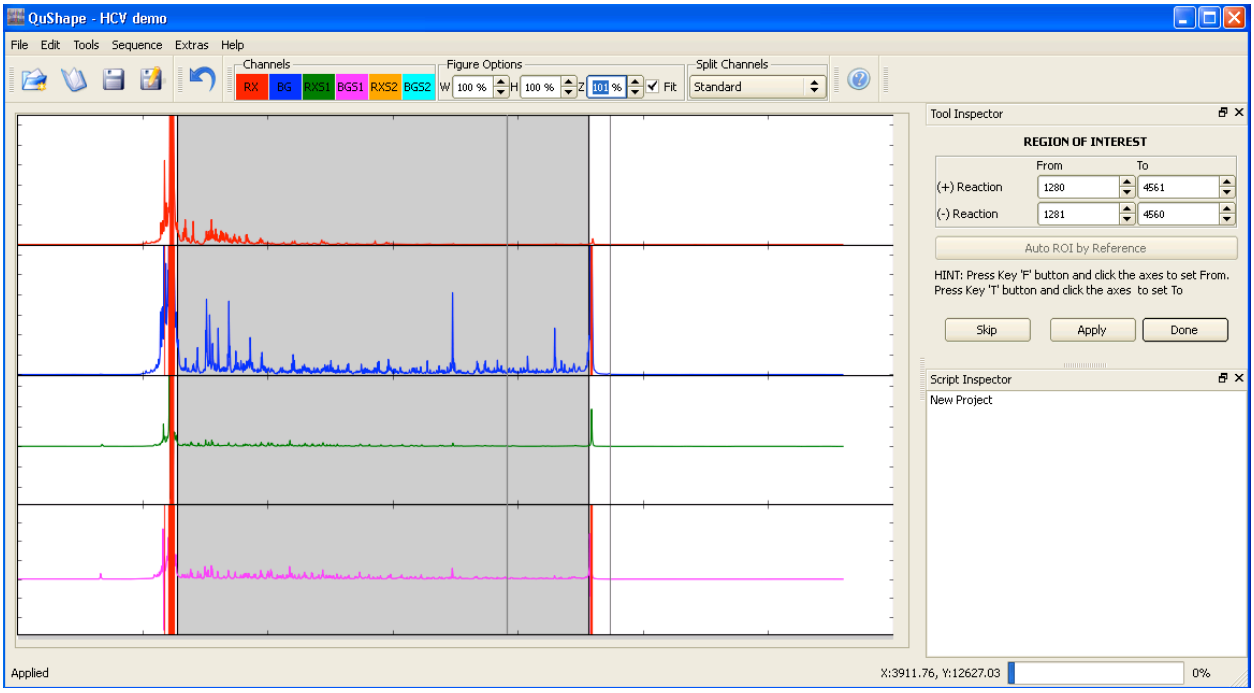
The pull-down **Tools** menu in the menu bar at the top of the screen has the following tool options: **Region of Interest**, **Smoothing**, **Mobility Shift**, **Baseline Adjustment**, **Signal Decay Correction**, and **Signal Alignment**. By default, these are executed by the program automatically in this order (and therefore there is no need for the user to open them manually, unless you wish to implement a particular tool out of this order). The standard procedure is for the user to execute each tool as it appears in the **Tool Inspector** window (by clicking the **Apply** button), inspect the result in the **Data View** window, and proceed to the next tool in the default sequence (by clicking the **Done** button). However, if you are confident that all the tools will perform satisfactorily and there is no need for visual inspection of the intermediate results, the **Apply All Tools** option can be selected from the **Tools** menu.



4.1. REGION OF INTEREST



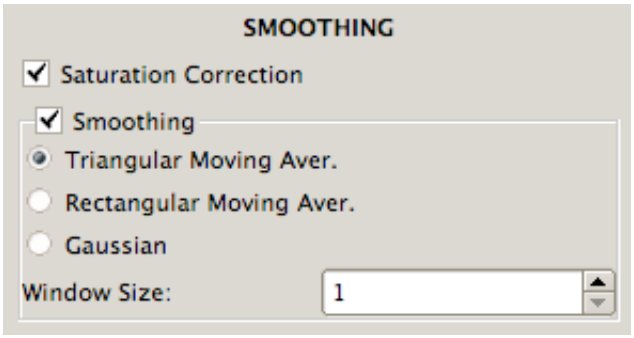
The start and the end segments of electropherogram traces typically have stretches of excessive and undifferentiated fluorescence that obscure peaks corresponding to the nucleotides at the either end of the studied RNA. The region of interest (ROI) must be selected along the elution time axis to avoid such unusable segments. This task is accomplished with the **Region of Interest** tool: Either type the elution time values of these points directly in the spin boxes in the **Tool Inspector** window or, more conveniently, select the start point of the ROI by pressing and holding down the 'F' (*from*) key on the keyboard and then placing the mouse arrow at the desired elution time position in the plot in the **Data View** window and clicking the left mouse button. The end point of the ROI is selected similarly by pressing and holding down the 'T' (*to*) key, and then placing the mouse arrow at the desired elution time position in the **Data View** window and clicking the left mouse button. Once the start and end points of the ROI are entered, the user-chosen ROI will be displayed in the **Data View** window on a gray background.



The ROI start and end points for the (+) reagent traces (i.e., RX, RXS1, and RXS2) are selected by typing their values in the **(+) Reaction** spin boxes or by pointing with the mouse in the RX panel. The ROI start and end points for the (–) reagent traces (i.e., BG, BGS1, and BGS2) are selected by typing their values in the **(–) Reaction** spin boxes or by pointing with the mouse in the BG panel. Note that in the **Data View** window the saturated portions of the traces are shaded in red.

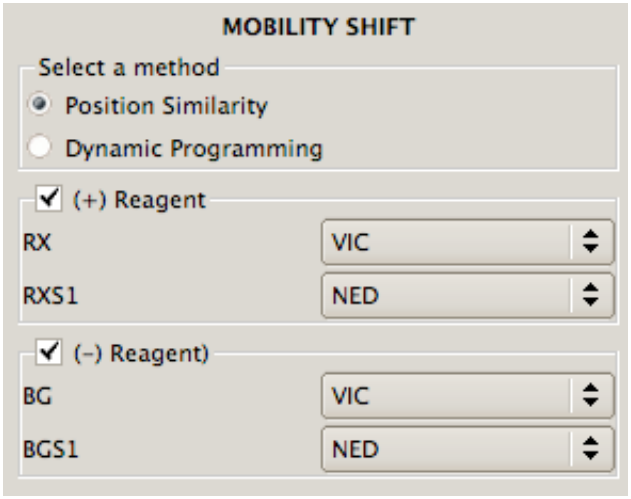
If the current data have a reference project, the **Auto ROI by Reference** button will be live in the **Tool Inspector** window. Clicking this button will find the ROI automatically by matching to the ROI in the reference project.

4.2. SMOOTHING



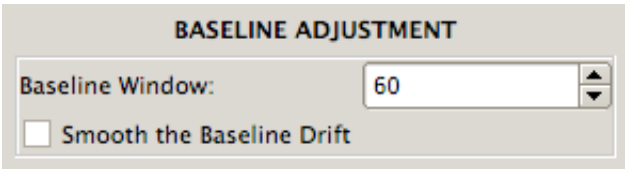
The **Smoothing** tool filters out high-frequency noise in the data and finds and corrects saturated data points. **Triangular Moving Average** is the default smoothing method. **Rectangular Moving Average** (boxcar) and **Gaussian** are also available. The width of these smoothing filters is shown in the **Window Size** box. The default value is 1; it can be increased if greater smoothing is desired.

4.3. MOBILITY SHIFT



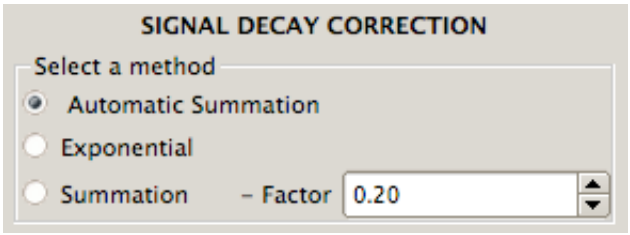
Use of different fluorescent labels in the same capillary results in slight differences in retention times for fragments of the same sequence and length. Therefore all data traces within each capillary have to be aligned relative to each other by time-shifting and time-scaling them along the elution time axis. The **Mobility Shift** tool performs this operation. The default mobility shift correction method is **Position Similarity**. A method based on dynamic time warping (**Dynamic Programming**) is also available.

4.4. BASELINE ADJUSTMENT



The **Baseline Adjustment** tool is used to remove baseline offset. Baseline is defined by the minima of a set of consecutive elution time intervals whose length is specified in the **Baseline Window** (default value is 60). If smoothing of the baseline drift is desired, the **Smooth the Baseline Drift** box should be checked.

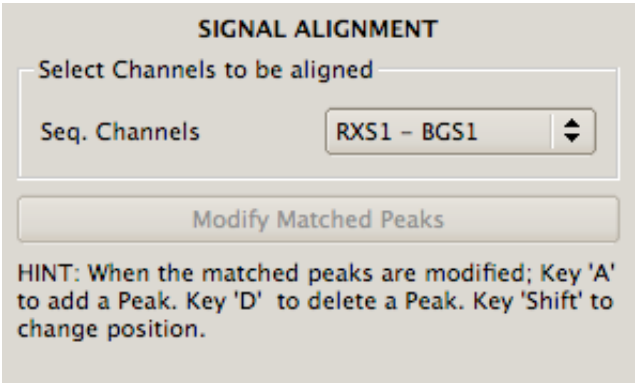
4.5. SIGNAL DECAY CORRECTION



A characteristic feature of fluorescent signals in SHAPE electropherograms is that intensity gradually declines as a function of the elution time due to technical reasons. Such gradual signal decay must be corrected computationally, and the **Signal Decay Correction** tool was designed to accomplish this task. The most effective method offered in this tool is **Automatic Summation**, and it is used as a default.

One alternative method is **Exponential**. It is an approximate method that relies on fitting an exponential function to the data and normalizing the data by this function. Another alternative method is **Summation**, which is the same method as Automatic Summation, but the user determines the value of the key parameter, **Factor**.

4.6. SIGNAL ALIGNMENT

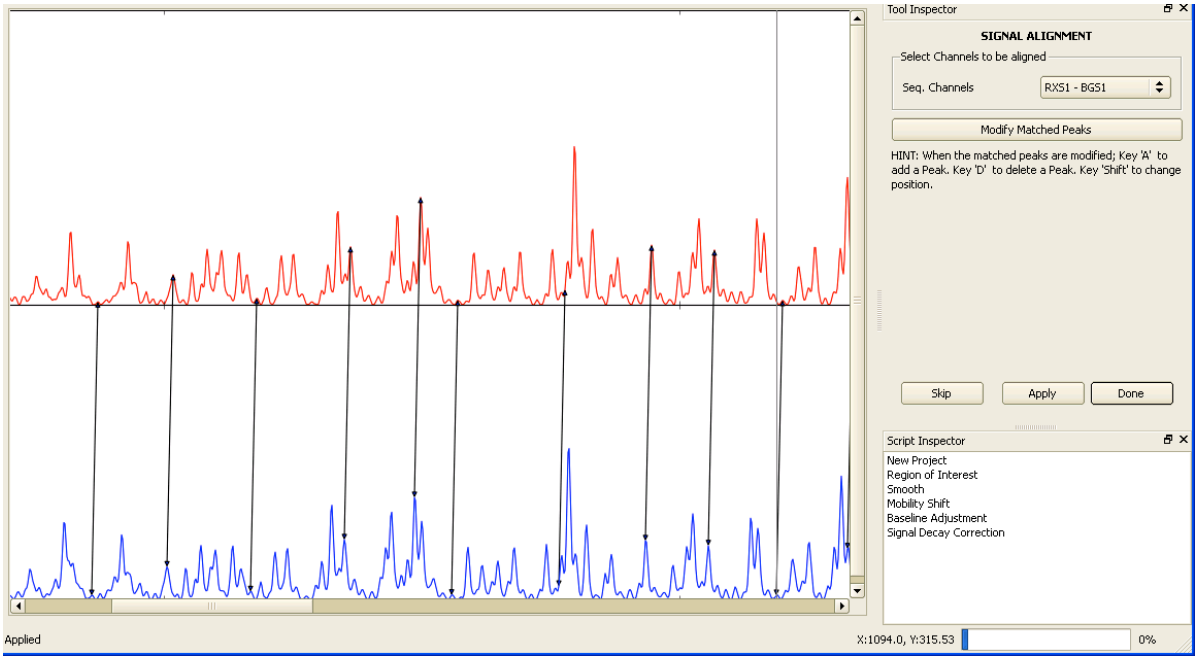


If (+) reagent and (–) reagent signals are measured in separate capillaries, the two traces must be aligned by time-shifting and time-scaling along the elution time axis. The **Signal Alignment** tool is used to align pairs of signals across two capillaries. Since signals within each

capillary are aligned using the **Mobility Shift** tool, it is necessary to perform the alignment of only two signals across two capillaries, and the other signals will become aligned as well. Because the sequencing signals in different capillaries (RXS1 and BGS1, and RXS2 and BGS2 if present) are much more similar to each other than are the reaction signals (RX and BG), the two capillaries are aligned using RXS1 and BGS1 signals (default option). If desired, RXS2 and BGS2 signals can be used in alignment instead by selection in the **Seq. Channels** window.

Signal alignment is a computationally intensive operation, typically taking several tens of seconds to complete. During this operation, the left-bottom corner of the screen will display the "**Applying…**" message. Once the operation is finished, this message will change to "**Applied**".

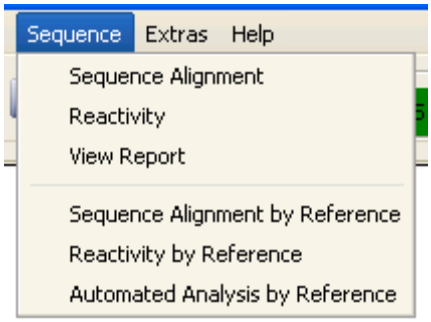
After the alignment procedure is finished, the aligned RX and BG signals will be plotted superimposed in one panel in the **Data View** window and the aligned RXS and BGS signals will be plotted superimposed in the other panel, so that the accuracy of the alignment can be checked visually. If misalignment is found, it can be corrected manually after clicking the **Modify Matched Peaks** button. This will change the display: RXS signal will be plotted above BGS signal and vertical lines will be drawn linking a subset of the matched peaks in the two signals.



An incorrect link between two peaks in the two signals can be changed by pressing and holding the '**Shift**' key, while placing the mouse arrow on the wrong peak, clicking and holding the left mouse button, and dragging the link to the desired peak. If a new link is desired, press and hold the '**A**' key while

clicking with the mouse on the two peaks that should be linked. If a link must be removed, press and hold the '**D**' key while clicking on that link with the mouse. Once all the desired link changes are made, click the **Apply** button to realign the two signals according to the newly imposed constraints.

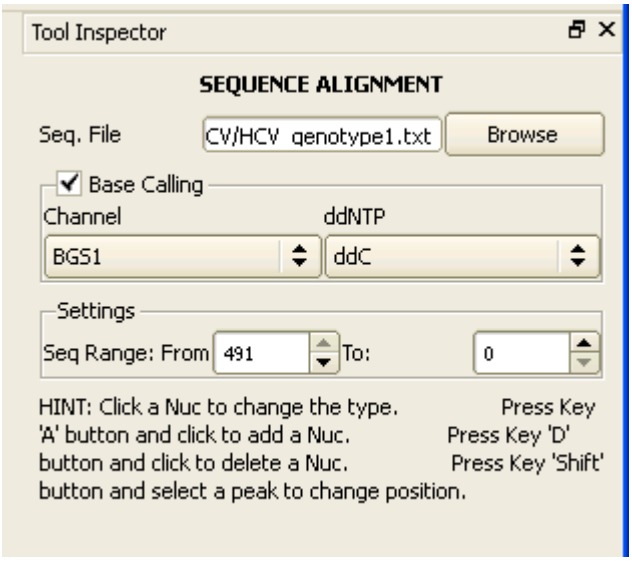
5. SEQUENCE MENU



The pull-down **Sequence** menu in the menu bar at the top of the screen has the following tool options that complete the analysis of the experimental data: **Sequence Alignment**, **Reactivity**, and **View Report**. By default, these tools are executed automatically and in this order after execution of the tools in the **Tools** menu (therefore there is no need for the user to open them manually, unless a particular tool is desired out of its default order). The **Sequence** menu also has a set of tools (**Sequence Alignment by Reference**, **Reactivity by Reference**, and **Automated Analysis by Reference**) that make use of a reference project, if available.

5.1. SEQUENCE ALIGNMENT

This tool performs three operations. First, it performs base calling, an operation that classifies all the peaks in the sequencing signal as either *specific* peaks produced by ddNTP-paired nucleotides or *non-specific* or background peaks corresponding to nucleotides of the other three bases. Next, the tool aligns peaks in the sequencing signal with the RNA nucleotide sequence. Finally, this tool assigns nucleotide-matched peaks in the sequencing signal to the corresponding peaks in the (+) reagent and (–) reagent signals, thus assigning each peak to its corresponding RNA position.

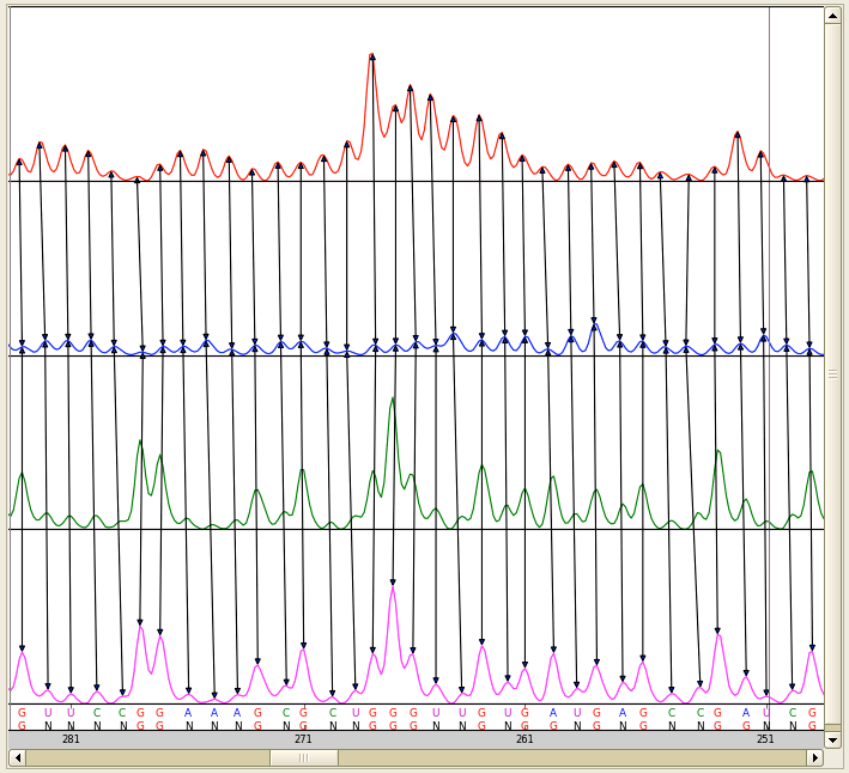


All three operations are performed automatically once the **Apply** button is clicked, relying on the information provided by the user during the steps involved in creating the **New** project. This information is displayed in five boxes in the **Tool**

Inspector window: RNA sequence file name (**Seq. File**); sequencing channel used for base calling and RNA alignment (**Channel**); dideoxynucleotide used (**ddNTP**); and the first and last nucleotides in the RNA sequence bracketing the studied RNA region (**Seq. Range: From … To**). If desired default settings can be changed to new ones by selection or entry in the appropriate boxes.

Of a particular practical importance, the default approach is to match the sequencing signal against the entire RNA sequence provided in the **Seq. File**. If the studied RNA section is much shorter than the entire RNA sequence in the file, however, the search can take a long time. In this case, narrow the search window by entering in the **Seq. Range: From … To** boxes the boundaries of the nucleotide sequence to be matched.

Sequence alignment is a computationally intensive operation, typically taking tens of seconds to complete. During this operation, the left-bottom corner of the screen will display the "**Applying…**" message. Once the operation has finished, this message will change to "**Applied**", and the display in the **Data View** window will change to a view in which corresponding peaks in RX, BG, and BGS traces are linked by vertical arrows. The results of base-calling and sequence alignment will be shown at the bottom of the BGS panel, with the top row showing the RNA sequence and the bottom row showing the results of base calling.



If the alignment is not accurate, the errors can be corrected manually. Four different manual correction operations are available:

(1) The base label of a peak in the BGS trace can be changed. For example, suppose that ddC was used for sequencing. Consequently, the bottom row consists of 'N' and 'G' labels. Clicking on 'N' with the mouse will turn it to 'G'. Clicking on 'G' will turn it to 'N'.
(2) An extra base can be added to the bottom row. By pressing and holding the '**A**' key while clicking at a particular location in the bottom row with the mouse, an 'N' will be inserted at that location and this added nucleotide will be linked to RX and BG.
(3) A base and corresponding links can be deleted by pressing and holding the '**D**' key while clicking at a base.
(4) Computed locations of the peak centers in BG and RX can be moved by pressing the '**Shift**' key and dragging the arrow to the desired location.

After modifying the sequences, press **Apply** to see the new alignment results with nucleotides matched to the peaks in RX and BG. Note that at this time the base calling operation will be disabled. If this operation needs to be performed again, check the **Base Calling** box.

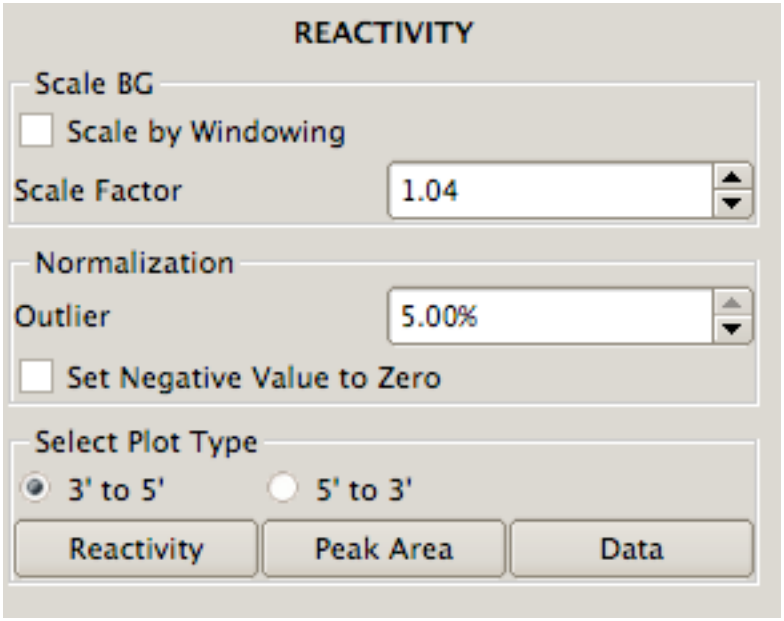
The **Base Calling** box is also important if the user wants to come back to the sequence alignment after pressing the '**Done**' button and moving to other tools. In that case, the **Sequence Alignment** tool can be called from the **Sequence** menu with the base-calling operation enabled, so that upon execution of this tool the previously manually corrected base assignments will be discarded. Therefore, the **Base Calling** box should be unchecked if you want to use previously obtained base-calling results.

5.2. REACTIVITY

This tool performs three operations:

First, a whole-signal Gaussian integration is performed for all peaks in the (+) and (–) reagent signals, fitting each peak with a Gaussian function individually optimized for position, height, and width.

Next, the scaling operation scales the BG signal relative to the RX signal. This scaling is necessary because the (+) and (–) reagent primer extension reactions are performed separately and not necessarily under fully identical conditions. When the **Reactivity** tool is open, the scaling factor is computed automatically and is displayed in the **Scale Factor** window. When the **Reactivity** tool is executed, by clicking the **Apply** button, the BG signal will be scaled by this factor. If not satisfied, other scaling factor values can be tested by entering them in the **Scale Factor** window.

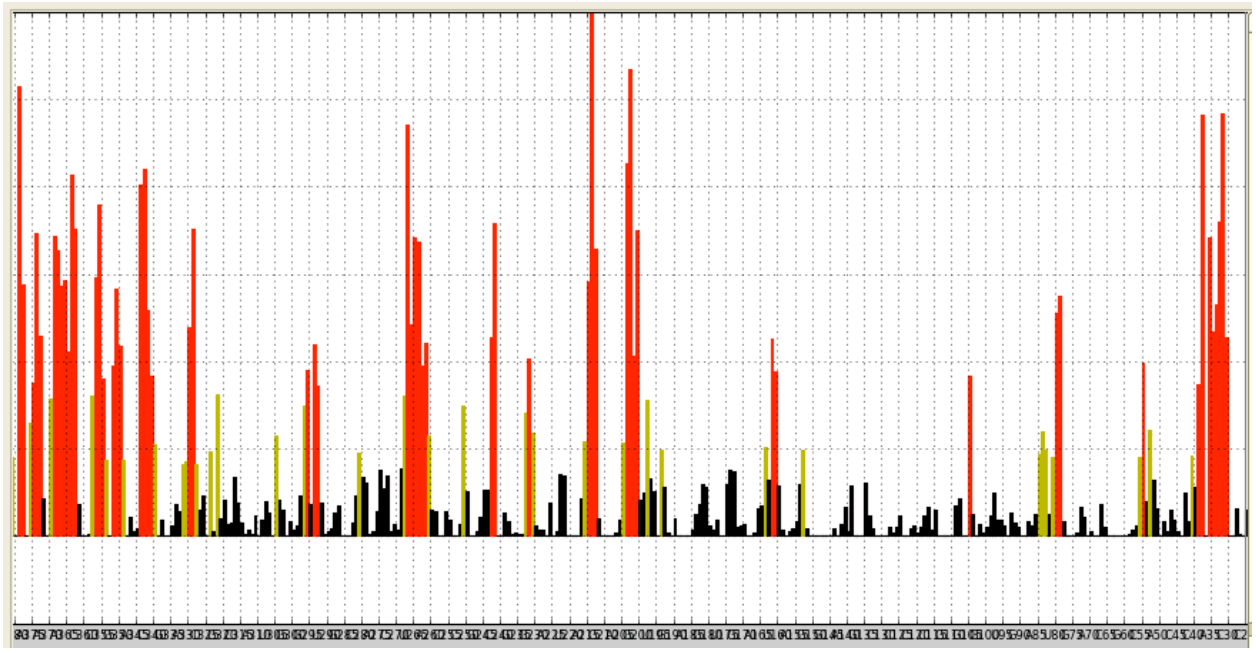


The '**Scale by Windowing**' box offers another user-controlled option. As a default, the scaling factor is automatically determined for the entire BG data set, and all BG signals are scaled by this factor. When working with very long sequences, it may be more accurate to scale BG locally, rather than globally. To use local scaling, check the '**Scale by Windowing**' box.

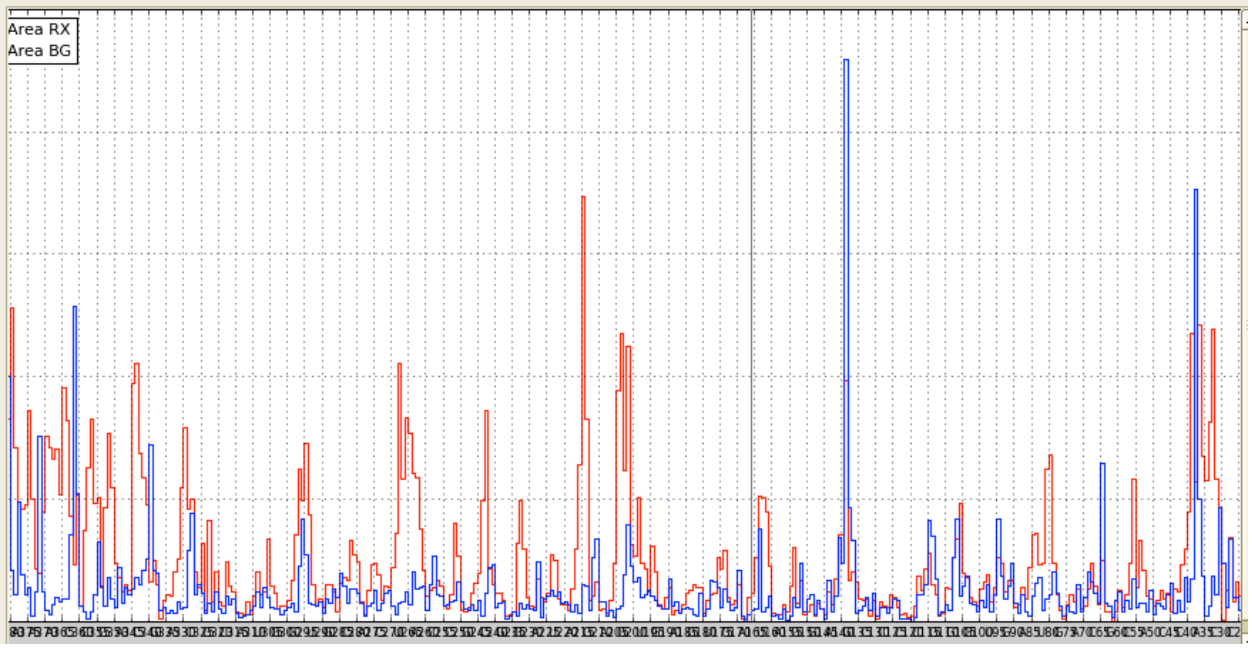
Finally, the normalization operation subtracts the integrated values for the (–) reagent peaks from the (+) reagent peaks, and normalizes the difference to obtain the normalized nucleotide-resolution reactivity for every RNA position. A box normalization-based algorithm is used to normalize data. This normalization scales reactivities to a scale spanning 0 to ~2, where zero indicates no reactivity and 1.0 is the average intensity for highly reactive RNA positions. Nucleotides with normalized SHAPE reactivities 0–0.4, 0.4–0.85, and >0.85 correspond to unreactive, moderately reactive, and highly reactive positions, respectively, and are plotted in different colors. As a part of the normalization procedure, the percent of outliers is determined automatically and is displayed in the **Outlier** window. A different percent of outliers can be selected.

There are three alternative displays of the output of the **Reactivity** tool:

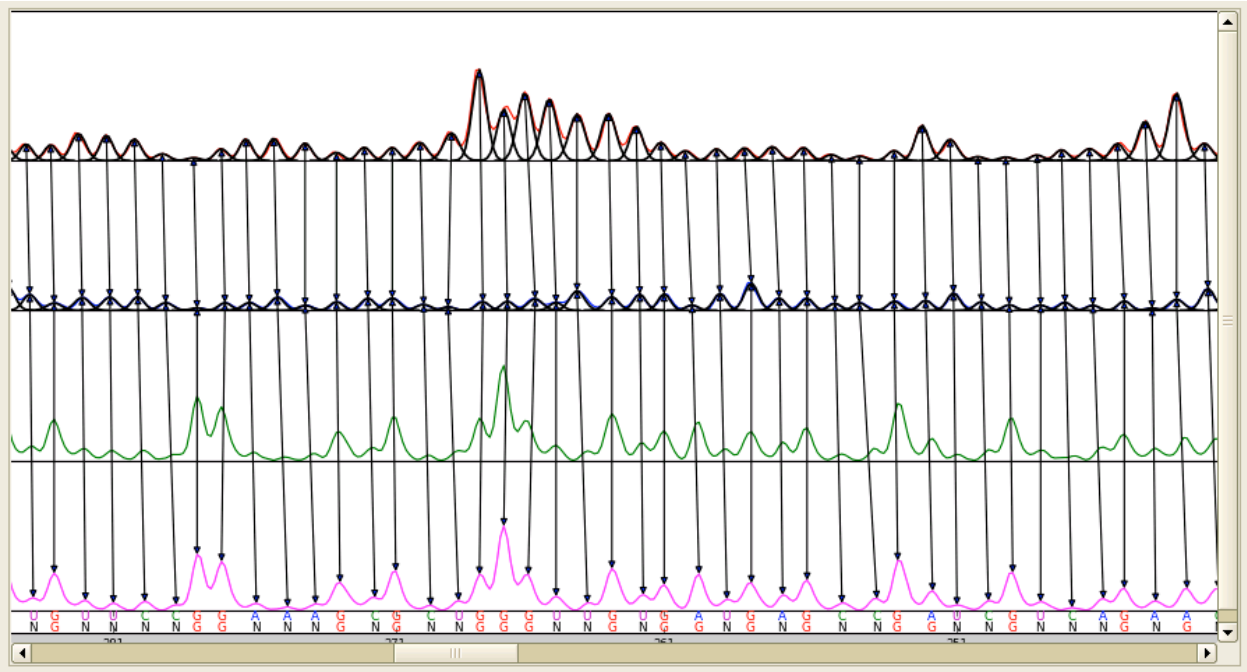
(1) "**Reactivity**" button plots the normalized reactivity of each nucleotide.



(2) "**Peak Area**" button plots the areas of RX and BG peaks.



(3) "**Data**" button draws the same plot as provided through the **Sequence Alignment** tool (linked RX, BG, RXS, and BGS traces, as well as the nucleotide sequence), but in addition it overlays each peak in RX and BG traces with its Gaussian estimation.



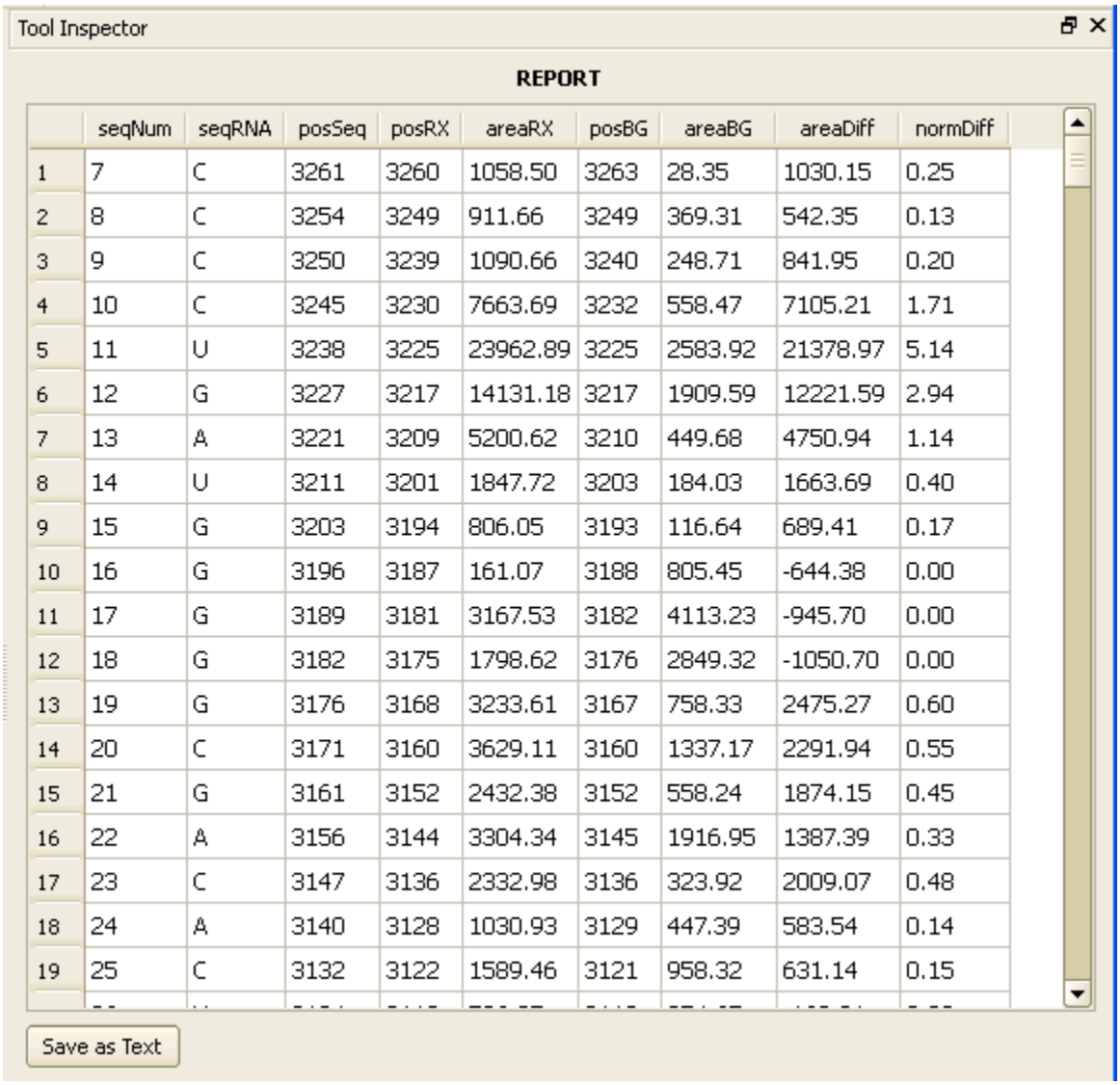
5.3. REPORT

The final output of QuShape data processing is a text file. This file contains information about each nucleotide including integrated (+) and (–) reagent peak areas (labeled RX and BG, respectively) and their subtracted, normalized SHAPE reactivities.

The final report of QuShape data processing is shown as a table in the **Tool Inspector** window. This table contains information about each nucleotide, including:

- **SeqNum**: Number of the nucleotide
- **seqRNA**: Nucleotide base type
- **posSeq**: Position of the nucleotide in the sequence ladder
- **posRX**: Position of the RX peak
- **areaRX**: Area of the RX peak
- **posBG**: Position of the BG peak
- **areaBG**: Area of the BG peak

- **areaDiff**: Difference between RX and BG areas
- **normDiff**: Normalized difference (the normalized reactivity of the nucleotide)



	seqNum	seqRNA	posSeq	posRX	areaRX	posBG	areaBG	areaDiff	normDiff
1	7	C	3261	3260	1058.50	3263	28.35	1030.15	0.25
2	8	C	3254	3249	911.66	3249	369.31	542.35	0.13
3	9	C	3250	3239	1090.66	3240	248.71	841.95	0.20
4	10	C	3245	3230	7663.69	3232	558.47	7105.21	1.71
5	11	U	3238	3225	23962.89	3225	2583.92	21378.97	5.14
6	12	G	3227	3217	14131.18	3217	1909.59	12221.59	2.94
7	13	A	3221	3209	5200.62	3210	449.68	4750.94	1.14
8	14	U	3211	3201	1847.72	3203	184.03	1663.69	0.40
9	15	G	3203	3194	806.05	3193	116.64	689.41	0.17
10	16	G	3196	3187	161.07	3188	805.45	-644.38	0.00
11	17	G	3189	3181	3167.53	3182	4113.23	-945.70	0.00
12	18	G	3182	3175	1798.62	3176	2849.32	-1050.70	0.00
13	19	G	3176	3168	3233.61	3167	758.33	2475.27	0.60
14	20	C	3171	3160	3629.11	3160	1337.17	2291.94	0.55
15	21	G	3161	3152	2432.38	3152	558.24	1874.15	0.45
16	22	A	3156	3144	3304.34	3145	1916.95	1387.39	0.33
17	23	C	3147	3136	2332.98	3136	323.92	2009.07	0.48
18	24	A	3140	3128	1030.93	3129	447.39	583.54	0.14
19	25	C	3132	3122	1589.46	3121	958.32	631.14	0.15

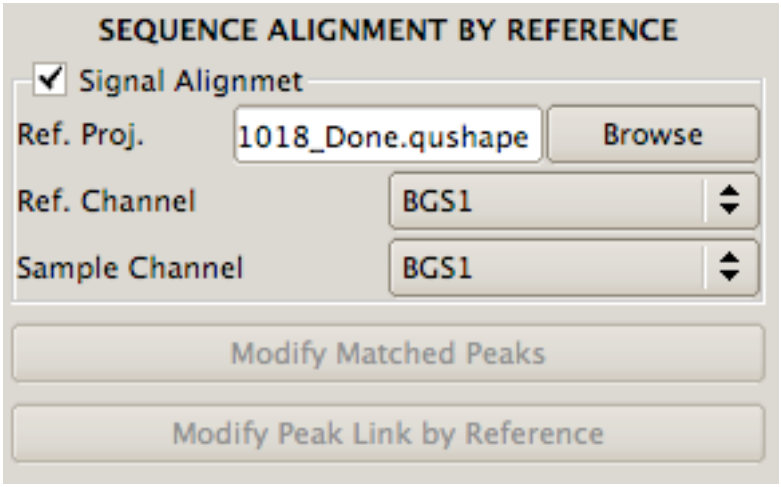
This table can be saved as a tab-delimited text file by clicking the '**Save as Text**' button.

5.4. SEQUENCE ALIGNMENT BY REFERENCE

QuShape allows the results of a previous analysis to serve as a reference for subsequent analyses on the same RNA. All parameters and the sequence alignment are saved in a "reference" project that can be used automatically in subsequent projects on the same RNA.

The **Sequence Alignment by Reference** tool uses signal and sequence alignments generated previously to align traces to each other and to the RNA sequence in the current project. The name of the reference file must be entered as a part of step 2 of creating the New Project (see **Section 3: New Project**, **step 2**). Alternatively, the name of the reference file can be entered in the **Tool Inspector** window opened with the **Sequence Alignment by Reference** tool by browsing **Ref. Proj**.

Default channels in the reference and current ('sample') projects that will be aligned are shown in the **Ref. Channel** and **Sample Channel** boxes. These channels can be changed if needed.



Clicking the **Apply** button will execute the alignment procedure and the outcome will be displayed in the same format as after execution of the **Sequence Alignment** tool (showing all four traces, with corresponding peaks linked by vertical lines, and displaying the RNA sequence at the bottom). In addition, the reference traces will be drawn (with light lines) superimposed on the current project traces for comparison.

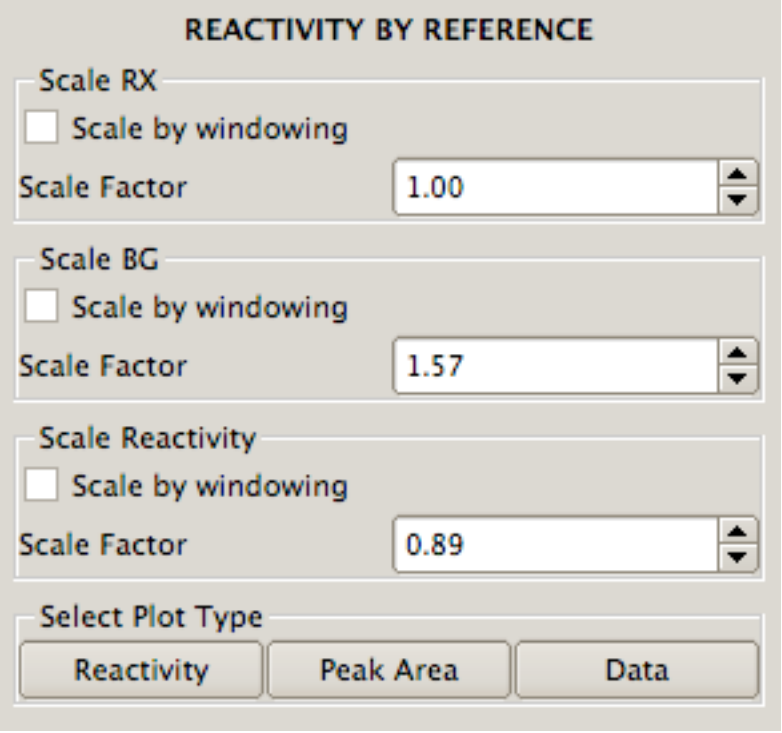
If there is any problem with the alignment, press the '**Modify Matched Peaks**' button to correct signal alignment manually. When this button is pressed, the selected reference channel (for example, Ref: BGS1) and sample channel (for example, Sample: BGS1) will be plotted with matched peaks. The matched peaks can then be modified using the same operations described in **Section 4.6: Signal Alignment**.

If the alignment between the two projects is accurate, but the sample peak has been linked with the incorrect reference peak, press the '**Modify Peak Link by Reference'** button**.** When this button is pressed, the selected "RX Reference" channel will be plotted with the "RX Sample", and the "BG

Reference" channel plotted with the "BG Sample". The peak link can then be modified for the Sample channels using the same operations described in **Section 4.6: Signal Alignment**.

If, while using the **Sequence Alignment by Reference** tool, a local error is found in the reference trace that requires that a peak be "added" or "deleted", the current sample project can be fixed easily. Press the **Done** button and use the pull-down **Sequence** menu in the menu bar at the top of the screen to select **Sequence Alignment**. Uncheck the **Base Calling** box before pressing the **Apply** button and the missing or extraneous peak can be added or deleted using the same operations described in **Section 5.1: Sequence Alignment**.

5.5. REACTIVITY BY REFERENCE



This tool fits Gaussian functions to each peak in the RX and BG traces and computes their areas. Sample RX and BG are then scaled to Reference RX and BG. Reactivity for each nucleotide is computed by subtracting BG from RX and then normalizing, as described in **Section 5.2: Reactivity**. The normalized reactivity can be scaled to the Reference normalized reactivity so that they accurately match the reference data. If the sample project used the **Sequence Alignment by Reference** tool, this is the default option. However, if desired, the **Reactivity** tool can be used by selecting it from the pull-down **Sequence** menu bar. For example, if there are large SHAPE reactivity differences between the reference and sample traces (as might be the case if different buffer conditions were used to anneal the RNA), it is better to use the usual **Reactivity** tool.